IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A SHORT REVIEW OF MULTI-MODAL SENTIMENT ANALYSIS

[1]Priya Singh,[2] VK Saraswat

[1]Department of Computer Science, Institute of Engineering and Technology, Dr. B. R. Ambedkar University, Khandari, Agra, India

[2]Department of Computer Science, Institute of Engineering and Technology, Dr. B. R. Ambedkar University, Khandari, Agra, India

Abstract: Multi-modal sentiment analysis (MSA) has emerged as a focal point of research efforts in recent years, primarily due to its capacity to utilize various data modalities—including text, audio, and visual content—to enhance the precision of sentiment predictions. This paper presents a concise review of recent progress in MSA, outlining pivotal methodologies, prominent challenges, and prospective research trajectories. The integration of deep learning frameworks, along with feature fusion techniques and the benchmark datasets commonly employed in MSA investigations, will be examined in detail. Furthermore, the discourse will extend to the challenges introduced by data heterogeneity, the issues surrounding synchronization, and the complexities related to interpretability in the context of MSA. The conclusion will propose several future directions, emphasizing the potential impact of large-scale pre-trained models and cross-modal learning approaches on the advancement of sentiment analysis research**.**

Index Terms - Multimodal, Fusion, Scalability, Explainabilty ,Modalities.

I. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a field of natural language processing (NLP) that aims to determine the sentiment expressed in a given piece of text, speech, or visual content. Traditional sentiment analysis has primarily relied on text-based methods; however, human emotions are often expressed through multiple modalities, such as facial expressions, speech intonations, and textual cues. Multi-modal sentiment analysis (MSA) seeks to incorporate these different data streams to improve sentiment prediction.

Recent advancements in deep learning and multi-modal fusion techniques have significantly enhanced MSA models, making them more effective in capturing complex emotional expressions. This review provides an overview of MSA approaches, challenges, and future research directions. The vision of sentiment analysis is first proposed by [1].

II. Key Methodologies in Multi-Modal Sentiment Analysis

The multimodal sentiment analysis has various methodologies to analyze the sentiments multimodal data representation, feature fusion techniques and deep learning approaches .The selection of methodologies depends on that what type of data a researcher uses .The aim of the methodologies is to identify the sentiments which is in the form of text- positive way, negative way and neutral way.

2.1 Multi-Modal Data Representation

The various modalities are available in multimodal representation of data where MSA leverages three primary modalities like text, audio and visual data. All data is gather from internet and social media sites for the sentiment analysis. Discussion over the modalities is as follows:

Text modality: Text sentiments are extracted from social media posts, reviews, or transcripts using NLP techniques. This extraction can be done by the various technologies like lexicon based, machine learning and deep learning.
Audio modality: Audio modality arises from speech synthesis(text to speech system) , speech recognition sound and music processing which captures tone, pitch, and speech rhythm to infer emotions.
Visual modality: visual modality generates from computer vision , image and video processing ,visual stories and visual attention which analyzes facial expressions, gestures, and micro-expressions.

2.2 Feature Fusion Techniques

The integration of multiple modalities is a crucial step in MSA. A multitype feature fusion technique is categorized in three categories like early fusion, late fusion and intermediate fusion. All the Common fusion methods are:

Early Fusion: is also called as Feature –level fusion it combines features from all modalities like text , audio or visual at the input level before feeding them into the model in to the single vector . The profit of feature-level fusion is that it allows for correlation among different multimodal features which would perform task completion in a better way. The challenge of early fusion is that the integration of different elements in this policy[9]. Another factor for this method is time synchronization of different modalities which gather from different areas of sentiments . Hence , the necessary requirement is before [processing the data , conversion of different modalities into a desired way. Fusion of modalities at the feature level poses the challenge of integrating widely dissimilar input features.

Late Fusion: is also called as Decision level Fusion . The separate modals are used to processes each modality separately and merges their predictions at the decision level. An issue in late fusion is that not all modalities capture communication between all, so its performance can be leverage and generates more complex interdependency between all modalities. Late Fusion is used to evaluate co-dependencies of modalities[9] .
In the realm of late fusion, each modality's features are processed and classified independently. Once classification is complete, these outcomes are integrated to form a final decision vector that facilitates sentiment prediction. This process is characterized as late fusion, as it involves the integration of results after the classification phase. Due to the complexities associated with early fusion, many researchers favor decision-level fusion. In this framework, each modality's input is modeled separately, and the results from individual modal analyses are synthesized at the end of the procedure. Within the domains of machine learning and pattern recognition, this strategy, often referred to as classifier fusion, is widely acknowledged for its efficacy in accommodating heterogeneous data types.
In the fields of machine learning and pattern recognition, this technique, often termed classifier fusion, is widely recognized for its effectiveness in handling diverse data types[9].

Intermediate Fusion: Intermediate fusion in sentiment analysis represents an advanced approach that integrates various sources and methodologies to enhance the complexity and accuracy of sentiment evaluation. Rather than relying on a singular model or dataset, this technique synthesizes multiple inputs, models, or features to yield a more comprehensive understanding of sentiment conveyed in textual data. By leveraging the strengths of diverse analytic frameworks, intermediate fusion facilitates a more nuanced interpretation of sentiments expressed, thereby contributing to improved outcomes in sentiment analysis research . The various deep learning models are used to do intermediate fusion like CNN , RNN. It utilizes deep learning models to extract modality-specific features before integrating them through attention mechanisms or feature transformers.

2.3 Deep Learning Approaches

Deep learning is a special type of ML(machine learning) first proposed in 1986.The techniques of deep learning are available to learn multiple levels of feature of abstraction to analyze data.

Convolutional Neural Networks (CNNs): CNN produces a higher accuracy and also trained the data speedily. And it requires large amount of train data sets and training time and applied for feature extraction from images and audio spectrograms. CNN consist of three layers of convolution with various filter sizes. The CNN takes input from word embedding and learns multilevel contextual features from each layer of CNN and implement a multilevel feature fusion [2].

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks: The word embedding shares to CNN similar share to and from here it learns temporal fusion in text data . RNN works for long distance relationship or communication of sequential data in text and speech as compared to other deep learning approaches it trained the data slowly.[3] .

Transformers and Attention Mechanisms: To enhance feature alignment across modalities, it is essential to learn contextual dependencies. The attention mechanism, a pivotal technique in this realm, was particularly examined in the context of image data, where it enables neural networks to concentrate on specific aspects of the information being processed. With the advent of high-performance concurrent computing, the application of the attention mechanism has become widespread in natural language processing tasks. Self-attention, a variant of the attention mechanism, plays a crucial role in optimizing computational resource utilization during the training of machine learning (ML) and deep learning (DL) models. This approach facilitates the model's capability to make predictions based on a data sample by extracting observations from various segments of the data. Consequently, implementing self-attention not only improves efficiency but also enhances the robustness of the predictive model.
Transformer [5]model is a type of deep learning model which helps to identify self –attention and is used to process and produces data in sequential manner. Transformer works on the scaled dot-production mechanism of attention .The multi-head module is a key component of transformer .While computing the attention , then it splits the inputs into smaller piece of data i.e. chunks and then implement dot-product on the received data. The transformer model follows architecture of encoder-decoder, which is known as NMT .

Graph Neural Networks (GNNs): IN GNN a model with pertained text with Glove Capture relationships between different modalities and enhance feature interaction. Graph Neural Networks (GNNs) represent a significant category of neural network architectures specifically designed for processing graph-structured data. This type of data encompasses a collection of entities, which are represented as nodes, and their interrelations, which are depicted as edges[7]. There are various types of GNNs, such as Recurrent Graph Neural Network(RGNN),HGNN (Heterogeneous graph Neural Network),Spatial Convolutional Neural Network, and Spectral Convolutional Neural Network the message-passing neural networks (MPNNs) have been introduced as a spatially-based graph filtering mechanism, enabling the effective aggregation of information across the graph structure.
The RGNN architecture is designed to facilitate scalable semi-supervised learning within the context of multi-relational graphs. This framework adeptly identifies intricate relationships and captures non-linear data correlations. Furthermore, it integrates these elements effectively and demonstrates robust scalability as a function of graph size, thereby ensuring enhanced performance outcomes.
The spatial convolutional network methodology draws inspiration from Convolutional Neural Networks (CNNs) and is specifically adapted for application within graph structures. Notably, spectral-based graph filters are grounded in robust mathematical principles derived from the theory of graph signal processing, providing a solid theoretical foundation for their effectiveness

Heterogeneous Graph Neural Networks (HGNNs) represent a significant advancement within the realm of Graph Neural Networks (GNNs), playing a pivotal role in the feature fusion of multimodal data by effectively mapping such data onto heterogeneous graph structures. This methodology facilitates the integration of manually defined rules with automated algorithms to derive meaningful meta-paths. The utilization of graph convolutional networks enables efficient propagation and aggregation of information across the graph. Moreover, the incorporation of virtual nodes enhances the capability to fuse information derived from various meta-paths. An attention mechanism is implemented to evaluate and compute the respective weights, allowing these virtual nodes to encapsulate the integration of multimodal datasets comprehensively. This innovative strategy not only expands the potential array of available meta-paths but also significantly improves the effectiveness of information fusion among long-distance related neighbors[6].

III. CHALLENGES IN MULTI-MODAL SENTIMENT ANALYSIS

Despite its advantages, MSA faces several challenges to enhance to productivity of data heterogeneity ,modality synchronization , interpretability of modality ,data scarcity and missing modality are the major challenges to handle the data in multimodal sentiment analysis.

3.1Data Heterogeneity: Data heterogeneity means the data from various domains and the varied modalities consist different structure, their representation and qualities [8]. For example, when modality from a time data another is a static image data ,both the modality differs in their tokens data noise , different distribution ,and the analysis of their different aspect how the data alignemtn and their performance will vary different modality for heterogeneous data either for the similar languages or for the different languages **.** Differences in data formats, sampling rates, and feature distributions across modalities.

3.2Synchronization Issues: Synchronization issues in multimodal sentiment analysis present significant challenges in processing and aligning data from various modalities, such as text, audio, video, and images, which may be received at differing times or rates. Effective synchronization is critical to ensuring that models accurately interpret the interrelationships among these modalities, for instance, the correlation between speech and facial expressions or between textual content and visual imagery. Inadequate handling of synchronization can result in inaccuracies within the analysis, ultimately leading to flawed sentiment predictions Aligning textual, audio, and visual data streams, especially in spontaneous conversations. The various category of synchronization are temporal and spatial .
Temporal synchronization means alignment of voice and video with each other. While spatial synchronization arises when speakers faces and voice is not correlate with each other.

3.3Interpretability: Interpretability holds significant importance in the realm of multimodal sentiment analysis due to the inherent complexity of processing and integrating diverse data modalities, each characterized by unique features, such as textual information, visual elements, and auditory signals. A thorough understanding of the interactions among these varying types of data and their contributions to sentiment predictions is essential for several critical reasons. These include fostering user trust, facilitating the debugging process, and enhancing the overall performance of the sentiment analysis model. Understanding how different modalities contribute to sentiment prediction in deep learning models.

3.4Data Scarcity: Data scarcity denotes the constrained accessibility of labeled or high-quality datasets essential for the effective training of machine learning models. Within the realm of multimodal sentiment analysis, this issue is particularly pronounced due to the requirement for systems to process and assimilate information across various modalities, including text, audio, video, and images. Each modality necessitates the construction of distinct labeled datasets, and the generation of extensive, high-quality datasets for multimodal learning frequently demands significant resources and incurs considerable costs. The limited availability of labeled multi-modal datasets hinders model training and generalization.

3.5Missing Modality: The phenomenon of missing modalities refers to the absence of specific modalities within a dataset, adversely affecting the efficacy of multimodal tasks. Traditional computational techniques typically employ autoencoders or generative adversarial networks to address the issue of missing modalities. However, these methods require substantial datasets and a commitment to data quality. Alternative methodologies, such as contrastive learning and meta-learning, have the potential to improve the generalization and stability of these computational approaches. Nonetheless, these strategies also necessitate the computation of generated modalities, which introduces an element of uncertainty into the process.

S.No.	Reference	To Given	To Do
1	C. Sindhu, et,al(2022)	It works on cross domain sentiment analysis**.**	Not able to show source and target domains. And over dependence on target domains.
2.	Ameen Abdullah Qaid Aqlan,(2019)	It shows the direction on big data techniques show better result in sentiment analysis.	Data acquisition is difficult for opinion mining **.**
3.	Yanying Mao (2024)	Only works for current challenges	Not focused on large training data models.
4.	Seunghyun Yoonet,al(2019)	It shows state of art methodology and produces result 68.8% to 71.8%.	Not focused on specific modalitiy.

5.	Doaa Mohey(2016)	Better discussion of effects of sentiments and evolution.	Does not show the effect of techniques and their responses.
6.	Ankita Gandhi ,et,al.(2023)	It shows the effectiveness of sentiment prediction and emotion recognition.	Varied modalities are not focused properly.
7.	RaziehAbedi Rad,et.al.(2023)	It shows the expression of power, flexibility, complex structure ,interpretability.	To do work on structured and unstructured data.

IV. FUTURE RESEARCH DIRECTIONS

The several promising future research directions can improve MSA modal using large scale pre trained models ,cross modal learning , explainabilty and fairness, robust and scalable systems are more vital use in future scope in multimodal as well as multilingual aspects.

4.1Large-Scale Pre-Trained Models: The large scale pre-trained models are first used in NLP. They are specially designed for the self –supervised learning and various network structures like Transformers which is mainly consisted of self-attention layers. Leveraging multi-modal transformers (e.g., CLIP, ViLBERT) for sentiment analysis. CLIP stands for contrastive language image pre-training which exhibits remarkable versatility during its pre training phase, successfully learning to execute a diverse array of tasks such as optical character recognition (OCR), geo-localization, and action recognition. Notably, it demonstrates superior performance compared to the leading publicly available ImageNet models, all while maintaining enhanced computational efficiency. Furthermore, our findings indicate that zero-shot CLIP models showcase significantly greater robustness than their supervised ImageNet counterparts, even when matched for accuracy [16].

ViBERT [extends the capabilities of the widely recognized BERT (Bidirectional Encoder Representations from Transformers) model, a frontrunner in the field of natural language processing (NLP). The key innovation presented by ViBERT lies in its incorporation of vision-based features within the BERT architecture, enabling it to process and analyze multimodal data, which includes both textual and visual information. By leveraging the complementary strengths of language models—predominantly for text comprehension—and vision models, specifically for image classification and recognition, ViBERT offers a holistic approach to understanding and interpreting complex datasets that encompass multiple modalities. A pre-trained visual model, known as the Vision Transformer, has been employed to enhance the analysis and understanding of visual modalities [11].

4.2Cross-Modal Learning: Cross-modal learning is an essential approach for enhancing information sharing across various modalities. This involves the integration of both global and local datasets, enabling the reconstruction of more informative sets through various self-supervised learning techniques. The implementation of cross-modal mechanisms effectively augments a model's capacity to capture not only the distinct contributions of individual modalities but also their interrelations. By employing dual-attention fusion and cross-modal feature mapping, this method facilitates a comprehensive integration of multimodal data, ensuring that sentiment information remains preserved across temporal sequences while maintaining consistency and robustness in the analysis. The dynamic nature of temporal data is crucial in modeling sequential information, and extensive research has focused on effectively capturing these temporal dependencies [10][11].

4.3 Explainability and Fairness: Explainability plays a crucial role in crafting interpretable models that enhance our understanding of multi-modal sentiment decision-making. In the framework of the EU General Data Protection Regulation (GDPR), the concept of explainability has become a significant legal challenge, especially concerning the "Right to explanation" for individuals impacted by automated decision-making systems. This situation demands a thorough grasp of the underlying mechanisms and learning processes of these models. To meet this requirement, model debugging is essential. It enables the scrutiny of both accurate and inaccurate predictions, informs design improvements, and identifies potential vulnerabilities, including artifacts in the datasets. This kind of analysis is vital for managing bias effectively and ensuring fairness in model results. Consequently, explainability emerges as a critical priority that warrants careful examination and action within the field of artificial intelligence systems[12].

The prevailing literature on machine learning fairness posits that biases inherent in machine learning models primarily from misrepresentations present in the training datasets. These biases are viewed as reflections of

existing societal biases, suggesting a direct correlation between the data utilized for training and the perpetuation of inequities in model performance [12]. Fair explainability focused on fair outcomes of sentiments.

4.4Robust and Scalable Systems: Robustness and scalability addressing real-time processing challenges for large-scale applications, such as customer feedback analysis and affective computing. Robustness is a critical attribute of a model, characterized by its capacity to effectively manage a diverse array of data variations and noise[19]. This capability is essential for ensuring consistent and accurate sentiment predictions across multiple contexts and data sources.

In dynamic, real-time environments such as live customer support, user sentiment can exhibit rapid fluctuations [20]. Therefore, it is imperative for a resilient Real-Time Sentiment Analysis (RMSA) system to effectively monitor and adapt to these variations without experiencing delays or inaccuracies. An inability to swiftly adjust to shifts in tone, mood, or context—exemplified by an abrupt change in user sentiment during an ongoing interaction—can result in erroneous sentiment predictions. Such inaccuracies are detrimental to the overall user experience, highlighting the necessity for continuous adaptation in sentiment analysis systems.

V. CONCLUSION

Multi-modal sentiment analysis presents an exciting avenue for improving emotion recognition by integrating textual, audio, and visual cues. While significant progress has been made in deep learning-based fusion techniques, challenges such as data heterogeneity, synchronization, and interpretability remain. Future research should focus on scalable, explainable, and robust models that can handle diverse and real-world sentiment analysis applications

REFRENCES

[1] Nasukawa Y (2003) Sentiment analysis: capturing favorability using natural language processing, IBM Almaden Research Center, CA 95120, https://doi.org/10.1145/945645.945658 .

[2] Yanying Mao , Qun Liu , Yu Zhang (april,2024) , Sentiment analysis methods, applications, and challenges: A systematic literature review, Published by Elsevier B.V. on behalf of King Saud University.

[3] Mohd usama ,wenjing xiao ,belal ahmad ,jiafu wan ,mohammad mehedi hassan,abdulhameed alelaiwi, Deep Learning based Weighted Feature Fusion Approach for Sentiment Analysis .Digital Object Identifier 10.1109/ACCESS.2019.DOI

[4] Lukas Stappen University of Augsburg Augsburg, Germany Alice Baird et,al. “The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress”. DOI: 10.48550/arXiv.2104.07123

[5]. Fanfei Meng*(corresponding), David Demeter, Sentiment analysis with adaptive multi-head attention in Transformer,2023.

[6] Peng, J.; He, Y.; Chang, Y.; Lu, Y.; Zhang, P.; Ou, Z.; Yu, Q. A Social Media Dataset and H-GNN-Based Contrastive Learning Scheme for Multimodal Sentiment Analysis. Appl. Sci. 2025, 15, 636. https://doi.org/10.3390/ app15020636.

[7] Razieh Abedi Rad ,Mohammad Reza Yamaghani ,Azamossadat Nourbakhsh,A survey of sentiment analysis methods based on graph neural network. DOI: https://doi.org/10.21203/rs.3.rs-3173515/v1,2023.

[8] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, 56(10): 1–42, 2024

[9]. Ankita Gandhi , Kinjal Adhvaryu , et.al,Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Available online 28 September 2022 1566-2535/© 2022 Elsevier B.V. All rights reserved.

[10]. Cai, Y., Li, X., Zhang, Y. *et al.* Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. *Sci Rep* **15**, 2126 (2025). https://doi.org/10.1038/s41598-025-85859-6.

[11]. Ning Ouyang , Enze Zhang ,et,al;Enhancing Multimodal Sentiment Analysis with Cross-Attention Enhanced Fusion Networks. DOI: https://doi.org/10.21203/rs.3.rs-5359494/v1.

[12]. Sma Balkır, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser, Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. DOI: 10.18653/v1/2022.trustnlp-1.8,2022.

[15] Xiao Wang1,2, Guangyao Chen1,3, Guangwu Qian1 , Pengcheng Gao1 , Xiao-Yong Wei1,4, Yaowei Wang()1 , Yonghong Tian()1,3 and Wen Gao1, Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey.2024.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748– 8763. PMLR, 2021.

[17] Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In 2016 IEEE 16th International Conference on Data Mining (ICDM) (pp. 439-448). doi:10.1109/icdm.2016.0055.

[18] K. Soni, P. Yadav, and Rahul, "Comparative Analysis of Rotten Tomatoes Movie Reviews using Sentiment Analysis," in 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1494-1500, doi: 10.1109/ICICCS53718.2022.9788287.

[19]. C. Sindhu, S. Thejaswin, S. Harikrishnaa and C. Kavitha, "Mapping Distinct Source and Target Domains on Amazon Product Customer Critiques with Cross Domain Sentiment Analysis," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 782-786, doi: 10.1109/ICAIS53314.2022.9742732.

[20] K. Chitra, A. Tamilarasi, S. G. Dharani, P. Keerthana and T. Madhumitha, "Opinion Mining and Sentiment Analysis on Product Reviews," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-7, doi: 10.1109/ICCCI54379.2022.9740777.